Supplementary to **An Efficient Bayesian Inference Framework for Coalescent-Based Nonparametric Phylodynamics**

Shiwei Lan[*], Julia A. Palacios[†], Michael Karcher[‡], Vladimir N. Minin[‡], and Babak Shahbaba[§]

May 21, 2015

1 SplitHMC

Here, we show how to solve Hamiltonian dynamics defined by (13) using the "splitting" strategy. Denote $\mathbf{z} := (\mathbf{f}, \mathbf{p})$, $\mathbf{V} := \mathbf{C}_{in}^{-1}\mathrm{e}^{\tau}$, and $\mathbf{A} := \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{V} & \mathbf{0} \end{bmatrix}$. The dynamics (14a) can be written as

$$\dot{\mathbf{z}} = \mathbf{A}\mathbf{z}. \tag{1}$$

We then have the analytical solution to (1) as

$$\mathbf{z}(t) = \mathrm{e}^{\mathbf{A}t}\mathbf{z}(0), \tag{2}$$

where $\mathrm{e}^{\mathbf{A}t}$ is a matrix defined as $\mathrm{e}^{\mathbf{A}t} := \sum_{i=0}^{\infty} \frac{t^i}{i!}\mathbf{A}^i$, which in turn can be written as follows:

$$\begin{aligned}
\mathrm{e}^{\mathbf{A}t} &= \begin{bmatrix} \mathbf{I} - \frac{t^2}{2!}\mathbf{V} + \frac{t^4}{4!}\mathbf{V}^2 + \cdots & \mathbf{I}t - \frac{t^3}{3!}\mathbf{V} + \frac{t^5}{5!}\mathbf{V}^2 + \cdots \\ -\mathbf{V}t + \frac{t^3}{3!}\mathbf{V}^2 - \frac{t^5}{5!}\mathbf{V}^3 + \cdots & \mathbf{I} - \frac{t^2}{2!}\mathbf{V} + \frac{t^4}{4!}\mathbf{V}^2 + \cdots \end{bmatrix} \\
&= \begin{bmatrix} \cos(\sqrt{\mathbf{V}}t) & \mathbf{V}^{-\frac{1}{2}}\sin(\sqrt{\mathbf{V}}t) \\ -\mathbf{V}^{\frac{1}{2}}\sin(\sqrt{\mathbf{V}}t) & \cos(\sqrt{\mathbf{V}}t) \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{V}^{-\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \cos(\sqrt{\mathbf{V}}t) & \sin(\sqrt{\mathbf{V}}t) \\ -\sin(\sqrt{\mathbf{V}}t) & \cos(\sqrt{\mathbf{V}}t) \end{bmatrix} \begin{bmatrix} \mathbf{V}^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}.
\end{aligned}$$

For positive definite matrix $\mathbf{V}$, we can use the spectral decomposition $\mathbf{V} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^{-1}$, where $\mathbf{U}$ is orthogonal matrix, i.e. $\mathbf{U}^{-1} = \mathbf{U}^{\mathsf{T}}$. Therefore we have

$$\mathrm{e}^{\mathbf{A}t} = \begin{bmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{0} & \mathbf{U} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}^{-\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \cos(\sqrt{\boldsymbol{\Sigma}}t) & \sin(\sqrt{\boldsymbol{\Sigma}}t) \\ -\sin(\sqrt{\boldsymbol{\Sigma}}t) & \cos(\sqrt{\boldsymbol{\Sigma}}t) \end{bmatrix} \cdot \begin{bmatrix} \boldsymbol{\Sigma}^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{U}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{U}^{-1} \end{bmatrix}. \tag{3}$$

In practice, we only need to diagonalize $\mathbf{C}_{in}^{-1}$ once: $\mathbf{C}_{in}^{-1} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1}$, then $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\mathrm{e}^{\tau}$. If we let $\mathbf{f}^* := \sqrt{\boldsymbol{\Lambda}}\mathrm{e}^{\tau/2}\mathbf{U}^{-1}\mathbf{f}$, $\mathbf{p}^*_{-D} := \mathbf{U}^{-1}\mathbf{p}_{-D}$, we have the solution (15) as follows

$$\begin{bmatrix} \mathbf{f}^*(t) \\ \mathbf{p}^*_{-D}(t) \end{bmatrix} = \begin{bmatrix} \cos(\sqrt{\boldsymbol{\Lambda}}\mathrm{e}^{\tau/2}t) & \sin(\sqrt{\boldsymbol{\Lambda}}\mathrm{e}^{\tau/2}t) \\ -\sin(\sqrt{\boldsymbol{\Lambda}}\mathrm{e}^{\tau/2}t) & \cos(\sqrt{\boldsymbol{\Lambda}}\mathrm{e}^{\tau/2}t) \end{bmatrix} \begin{bmatrix} \mathbf{f}^*(0) \\ \mathbf{p}^*_{-D}(0) \end{bmatrix}.$$

We then apply leapfrog method to the remaining dynamics.

[*]Department of Statistics, University of Warwick, Coventry CV4 7AL.
[†]Department of Organismic and Evolutionary Biology, Harvard University.
[‡]Departmetn of Statistics, University of Washington
[§]Department of Statistics, University of California, Irvine.

2 Effect of precision matrix on efficiency of ES^2

For all experiments in the paper, we use the intrinsic precision matrix $\mathbf{C}_{in}^{-1}$ with small number added to its $(1,1)$ element in order to make $\mathbf{C}_{in}^{-1}$ invertible for ES^2. Since ES^2 is designed explore the dominant Gaussian component of posterior, it is inevitably affected by the structure Gaussian prior. We found indeed that the sampling efficiency depend on the precision matrix of Gaussian process in various example.

We implement ES^2 with the following 3 type of precision matrices:

- precision matrix (tri-diagonal) of a Brownian motion, $\mathbf{C}_{BM}^{-1}$, denoted as 'bmC'.
- intrinsic precision matrix $\mathbf{C}_{in}^{-1}$ with nugget to the whole diagonal elements, denote as 'diagnugget'.
- intrinsic precision matrix $\mathbf{C}_{in}^{-1}$ with nugget at (1,1) element (used in the paper), denoted as '11nugget'.

We compare sampling efficiency of ES^2 under 3 different settings in Tables S1 and S2.

simulation	setting	cond	s/iter	Eff($\mathbf{f}$)	minEff($\mathbf{f}$)/s	Eff(τ)	Eff(τ)/s
logistic	bmC	7.01E+04	1.46E-03	(35,88,180)	1.60	12.95	0.59
	diagnugget	1.48E+06	1.56E-03	(26,56,174)	1.13	8.86	0.38
	11nugget	1.80E+08	1.62E-03	(5,15,50)	0.19	6.58	0.27
exp	bmC	8.39E+04	1.52E-03	(21,64,145)	0.91	16.38	0.72
	diagnugget	5.14E+06	1.54E-03	(36,69,118)	1.55	10.60	0.46
	11nugget	5.48E+08	1.68E-03	(6,14,58)	0.22	7.14	0.28
boombust	bmC	8.09E+04	1.57E-03	(20,52,119)	0.84	12.98	0.55
	diagnugget	4.13E+06	1.53E-03	(33,67,122)	1.45	8.90	0.39
	11nugget	4.48E+08	1.67E-03	(5,14,38)	0.21	8.23	0.33
bottleneck	bmC	6.79E+04	1.43E-03	(81,169,435)	3.78	24.44	1.14
	diagnugget	1.32E+07	1.57E-03	(15,34,75)	0.66	7.15	0.31
	11nugget	1.35E+09	1.66E-03	(6,16,83)	0.25	3.47	0.14

Table S1: Sampling Efficiency of ES^2 in modeling simulated population trajectories. cond is the condition number (ratio of the largest and the smallest eigenvalues) of the precision matrix. s/iter is the seconds per sampling iteration. Eff includes the (min,med, max) of the effective sample size, and minEff/s is the time normalized the effective sample size.

setting	cond	s/iter	Eff($\mathbf{f}$)	minEff($\mathbf{f}$)/s	Eff(τ)	Eff(τ)/s
bmC	1.10E+05	1.74E-03	(5,25,99)	0.20	11.00	0.42
diagnugget	4.20E+04	1.76E-03	(9,30,111)	0.32	7.86	0.30
11nugget	2.01E+07	1.88E-03	(4,15,98)	0.15	15.89	0.57

Table S2: Sampling Efficiency of ES^2 in Influenza problem

As one can see that in general the larger condition number (ratio of the largest and the smalles eigenvalues) is, the lower sampling efficiency (measured by minEff/s) ES^2 will have. This is because the more ill-conditioned the precision matrix is, the higher autocorrelated are these states traversed by ES^2 on the ellipsis. Our proposed splitHMC performs uniformly well with different precision matrices without suffering from ill condition.

3 Simulation of four genealogies

Our four simulated genealogies are displayed in Figure S1. To simulate our data under heterochronous sampling, we selected 10 of our samples to have sampling time 0 and the rest of our 40 samples had sampling times assigned uniformly at random. Our four simulated genealogies were generated in `R` and our code is available online.

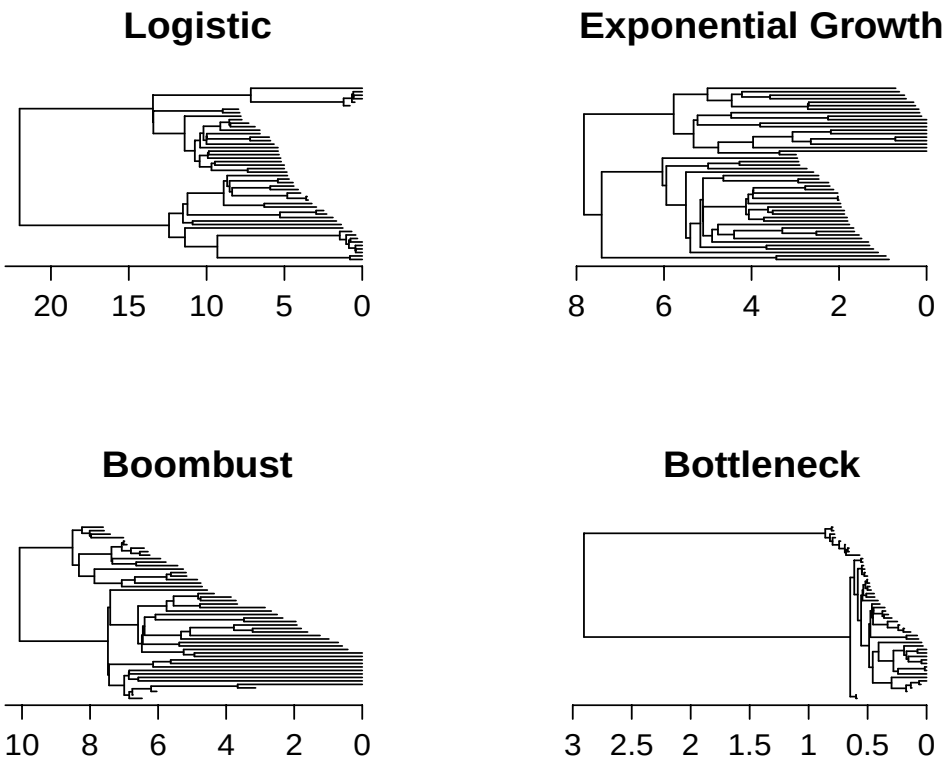

Figure S1: Simulated genealogies of I) logistic, II) exponential growth, III) boombust and IV) bottleneck population size trajectories.

4 Comparison with other methods

Figure 4 shows the estimated population size trajectory for the four simulations using our splitHMC, Bayesian Skyline Plot (Drummond et al., 2005) and Bayesian Skyride (Minin et al., 2008). In Figure S2, we show the trace plots of the posterior distributions of the results displayed in Figure 4 to asses convergence of the posterior estimates.

5 Effect of grid size on estimation accuracy by splitHMC

To study the effect of grid size on the accuracy of inference, we implement the proposed splitHMC algorithm on the bottleneck example. Population size trajectories are estimated with posterior samples generated by splitHMC using 20, 50, 100 and 200 grid points respectively. As seen in Figure S3, these estimates are quite close to each other given sufficient grid points, e.g. 50 points in this example, and their accuracy is not sensitive to the grid size.

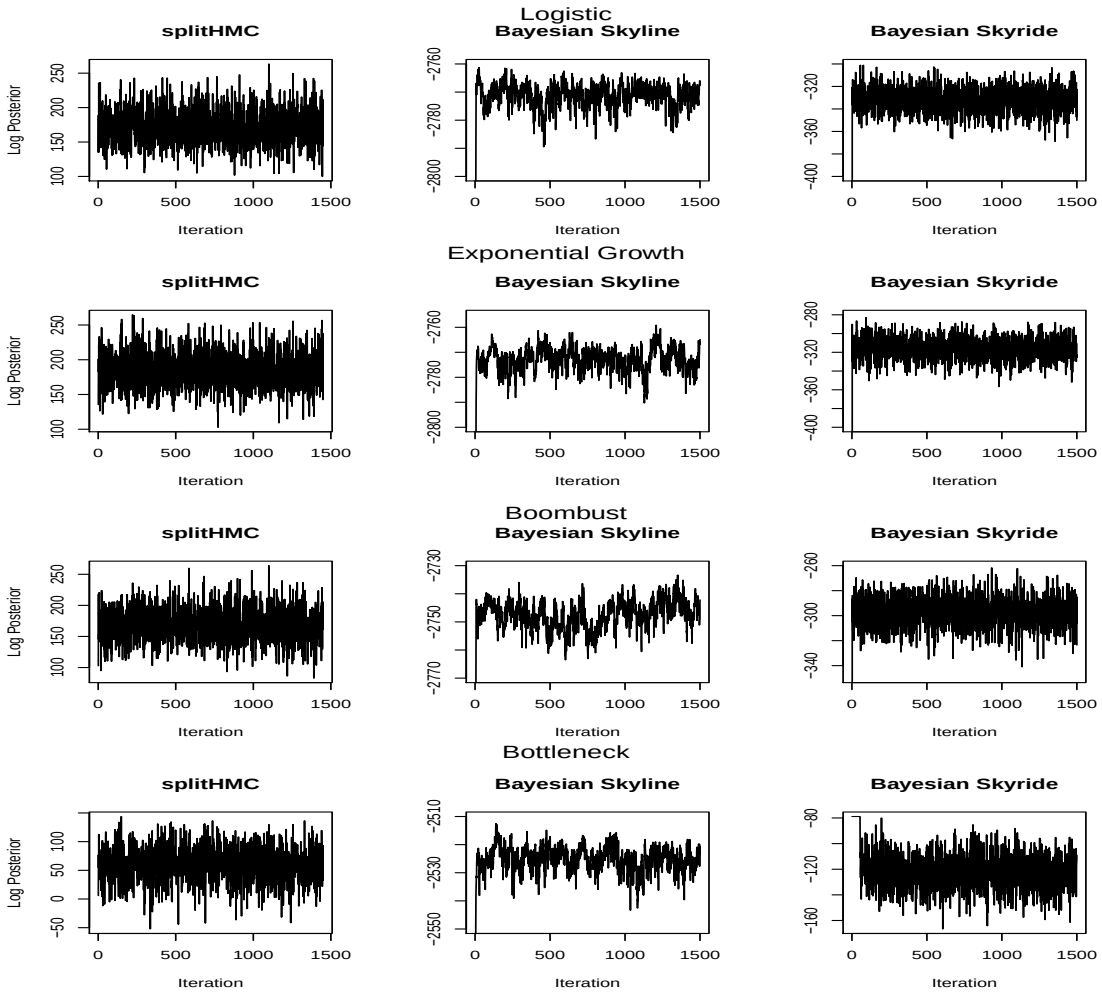

Figure S2: Trace plots of log posterior distributions of results displayed in Figure 4.

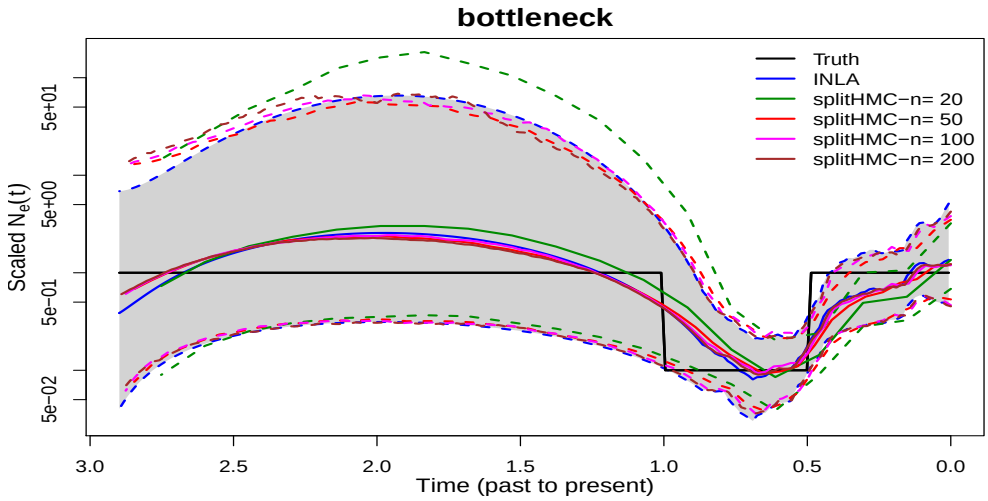

Figure S3: INLA vs splitHMC in estimating bottleneck population size trajectory. Shaded region shows 95% credible interval given by INLA and dotted lines show 95% credible intervals estimated with MCMC samples given by splitHMC.

6 Eigen decomposition vs Cholesky decomposition in splitHMC

For our splitHMC method, we use the eigen decomposition of $\mathbf{C}_{in}^{-1}$ with $\mathcal{O}(D^2)$ computational complexity. As discussed in the paper, we could instead use the Cholesky decomposition, $\mathbf{C}_{in}^{-1} = \mathbf{R}^\mathsf{T}\mathbf{R}$, to reduce the complexity to $\mathcal{O}(D)$. To this end, we could use change of variables $\mathbf{f}^* = \mathbf{R}\mathbf{f}$, which results in a simpler Hamiltonian dynamics compared to (14a) (Pakman and Paninski, 2014). Denote $\boldsymbol{\theta}^* := (\mathbf{f}^*, \tau)$. Then, the Hamiltonian in equation (13) can be rewritten as

$$H(\boldsymbol{\theta}^*, \mathbf{p}) = \frac{-l - [(D-1)/2 + \alpha]\tau + \beta e^\tau}{2} + \frac{(\mathbf{f}^*)^\mathsf{T}\mathbf{f}^* e^\tau + \mathbf{p}^\mathsf{T}\mathbf{p}}{2} + \frac{-l - [(D-1)/2 + \alpha]\tau + \beta e^\tau}{2}. \quad (4)$$

Note that the resulting Hamiltonian is defined in terms of $\mathbf{f}^*$ and τ. Consequently, the dynamics defined by the middle term of Equation (4) can be split as follows:

$$\begin{cases} \dot{\mathbf{f}}^* | \tau = \mathbf{p}_{-D}, \\ \dot{\mathbf{p}}_{-D} = -\mathbf{f}^* e^\tau. \end{cases} \quad (5a) \qquad \begin{cases} \dot{\tau} | \mathbf{f}^* = p_D, \\ \dot{p}_D = -(\mathbf{f}^*)^\mathsf{T}\mathbf{f}^* e^\tau / 2, \end{cases} \quad (5b)$$

The solution of equation (5a) is

$$\begin{aligned} \mathbf{f}^*(t) &= \mathbf{f}^*(0)\cos(e^{\tau/2}t) + \mathbf{p}_{-D}(0)e^{-\tau/2}\sin(e^{\tau/2}t) \\ \mathbf{p}_{-D}(t) &= -\mathbf{f}^*(0)e^{\tau/2}\sin(e^{\tau/2}t) + \mathbf{p}_{-D}(0)\cos(e^{\tau/2}t) \end{aligned} \quad (6)$$

Algorithm 1 can be easily modified according to this alternative dynamics. Because $\mathbf{R}$ is tridiagonal, backward solving $\mathbf{f}$ from $\mathbf{f}^*$ takes $\mathcal{O}(D)$ operations. Therefore, in theory the overall computational complexity could be reduced to $\mathcal{O}(D)$. In practice, however, we found that this approach would work well in cases where the middle term in Equation (4) dominates the Hamiltonian. In such cases, the residual energy, i.e., the first and last term in equation (4), is negligible so scaling its corresponding gradient by the Cholesky factor obtained from the middle part (as required by the change of variable),

$$\mathbf{s}^* := \nabla_{\mathbf{f}^*} l(\mathbf{f}) = \frac{d\mathbf{f}^\mathsf{T}}{d\mathbf{f}^*}\nabla_{\mathbf{f}} l(\mathbf{f}) = \mathbf{R}^{-\mathsf{T}}\mathbf{s}, \quad \text{where } \mathbf{s} := \nabla_{\mathbf{f}} l(\mathbf{f}) \quad (7)$$

will not have a substantial effect on the overall trajectory. Because this condition does not hold for examples discussed in this paper, we found that using the Cholesky decomposition would lead to low acceptance rates since the corresponding approximation error for simulating trajectories would be relatively high. To increase acceptance rate, we could reduce the stepsize, but this would negatively affect the overall efficiency of our splitHMC method. Therefore, as discussed in the paper, we used the eigen decomposition instead.

7 Adaptive MALA algorithm

We now show that the joint block updating in Knorr-Held and Rue (2002) can be recognized as an adaptive MALA algorithm. First, we sample $\kappa^* | \kappa \sim p(\kappa^* | \kappa) \propto \frac{\kappa^* + \kappa}{\kappa^* \kappa}$ on $[\kappa/c, \kappa c]$ for some $c > 1$ controlling the step size of κ. Denote $\mathbf{w} := \{A_{i,k}\Delta_\alpha\}_1^{D+m+n-4}$ and use the following Taylor expansion for $\log p(\mathbf{f}|\kappa)$ about $\hat{\mathbf{f}}$:

$$\begin{aligned} \log p(\mathbf{f}|\kappa) &= -\mathbf{y}^\mathsf{T}\mathbf{f} - \mathbf{w}^\mathsf{T}\exp(-\mathbf{f}) - \frac{1}{2}\mathbf{f}^\mathsf{T}\kappa\mathbf{C}_{in}^{-1}\mathbf{f} \\ &\approx -\mathbf{y}^\mathsf{T}\mathbf{f} - (\mathbf{w}\exp(-\hat{\mathbf{f}}))^\mathsf{T}[\mathbf{1} - (\mathbf{f} - \hat{\mathbf{f}}) + (\mathbf{f} - \hat{\mathbf{f}})^2/2] - \frac{1}{2}\mathbf{f}^\mathsf{T}\kappa\mathbf{C}_{in}^{-1}\mathbf{f} \\ &= (-\mathbf{y} + \mathbf{w}\exp(-\hat{\mathbf{f}})(\mathbf{1} + \hat{\mathbf{f}}))^\mathsf{T}\mathbf{f} + \frac{1}{2}\mathbf{f}^\mathsf{T}[\kappa\mathbf{C}_{in}^{-1} + \mathrm{diag}(\mathbf{w}\exp(-\hat{\mathbf{f}}))]\mathbf{f} \\ &=: \mathbf{b}^\mathsf{T}\mathbf{f} - \frac{1}{2}\mathbf{f}^\mathsf{T}\mathbf{G}\mathbf{f} = -\frac{1}{2}(\mathbf{f} - \mathbf{G}^{-1}\mathbf{b})^\mathsf{T}\mathbf{G}(\mathbf{f} - \mathbf{G}^{-1}\mathbf{b}) + \text{const} \end{aligned}$$

where $\mathbf{b}(\hat{\mathbf{f}}) := -\mathbf{y} + \mathbf{w}\exp(-\hat{\mathbf{f}})(\mathbf{1} + \hat{\mathbf{f}})$, $\mathbf{G}(\hat{\mathbf{f}}, \kappa) := \kappa\mathbf{C}_{in}^{-1} + \mathrm{diag}(\mathbf{w}\exp(-\hat{\mathbf{f}}))$. Setting $\hat{\mathbf{f}}$ to the current state, $\mathbf{f}$, and propose $\mathbf{f}^* | \mathbf{f}, \kappa^*$ from the following Gaussian distribution:

$$\begin{aligned} \mathbf{f}^* | \mathbf{f}, \kappa^* &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ \boldsymbol{\mu} &= \mathbf{G}(\mathbf{f}, \kappa^*)^{-1}\mathbf{b}(\mathbf{f}) = \mathbf{f} + \mathbf{G}(\mathbf{f}, \kappa^*)^{-1}\nabla_{\mathbf{f}} \log p(\mathbf{f}|\kappa^*) \\ \boldsymbol{\Sigma} &= \mathbf{G}(\mathbf{f}, \kappa^*)^{-1} \end{aligned}$$

Algorithm 1 Adaptive MALA (aMALA)

Given current state $\boldsymbol{\theta} = (\mathbf{f}, \kappa)$ calculate potential energy $U(\boldsymbol{\theta})$
repeat
  $z \sim \text{Unif}[1/c, c],\ u \sim \text{Unif}[0, 1]$
until $u < \frac{z+1/z}{c+1/c}$
update precision parameter $\kappa^* = \kappa z$
Sample momentum $\mathbf{p} \sim \mathcal{N}(\mathbf{0}, \mathbf{G}(\mathbf{f}, \kappa^*)^{-1})$
Calculate log of proposal density $\log p(\mathbf{f}^*|\mathbf{f}, \kappa^*) = -\frac{1}{2}\mathbf{p}^\mathsf{T}\mathbf{G}(\mathbf{f}, \kappa^*)\mathbf{p} + \frac{1}{2}\log\det\mathbf{G}(\mathbf{f}, \kappa^*)$
update momentum $\mathbf{p} \leftarrow \mathbf{p} - \varepsilon/2\mathbf{G}(\mathbf{f}, \kappa^*)^{-1}\nabla U(\mathbf{f}, \kappa^*)$
update latent variables $\mathbf{f}^* = \mathbf{f} + \varepsilon\mathbf{p}$
update momentum $\mathbf{p} \leftarrow \mathbf{p} - \varepsilon/2\mathbf{G}(\mathbf{f}^*, \kappa)^{-1}\nabla U(\mathbf{f}^*, \kappa)$
Calculate log of reverse proposal density $\log p(\mathbf{f}|\mathbf{f}^*, \kappa) = -\frac{1}{2}\mathbf{p}^\mathsf{T}\mathbf{G}(\mathbf{f}^*, \kappa)\mathbf{p} + \frac{1}{2}\log\det\mathbf{G}(\mathbf{f}^*, \kappa)$
Calculate new potential energy $U(\boldsymbol{\theta}^*)$
Accept/reject the proposal according to $\log\alpha = -U(\boldsymbol{\theta}^*) + U(\boldsymbol{\theta}) - \log p(\mathbf{f}^*|\mathbf{f}, \kappa) + \log p(\mathbf{f}|\mathbf{f}^*, \kappa)$ for the next state $\boldsymbol{\theta}'$

which has the same form as Langevin dynamical proposals. Interestingly, $\mathbf{G}(\hat{\mathbf{f}}, \kappa)$ is exactly the (observed) Fisher information. That is, this approach is equivalent to Riemannian MALA (Girolami and Calderhead, 2011).

Finally, $\boldsymbol{\theta}^* = (\mathbf{f}^*, \kappa^*)$ is jointly accepted with the following probability:

$$\alpha = \min\left\{1, \frac{p(\boldsymbol{\theta}^*|\text{data})}{p(\boldsymbol{\theta}|\text{data})}\frac{p(\kappa|\kappa^*)p(\mathbf{f}|\mathbf{f}^*, \kappa)}{p(\kappa^*|\kappa)p(\mathbf{f}^*|\mathbf{f}, \kappa^*)}\right\} = \min\left\{1, \frac{p(\boldsymbol{\theta}^*|\text{data})}{p(\boldsymbol{\theta}|\text{data})}\frac{p(\mathbf{f}|\mathbf{f}^*, \kappa)}{p(\mathbf{f}^*|\mathbf{f}, \kappa^*)}\right\}$$

where $p(\kappa^*|\kappa)$ is a symmetric proposal. The steps of adaptive MALA are summarized in algorithm (1).

References

Drummond, A., Suchard, M., Xie, D., and Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29:1969–1973.

Drummond, A. J., Rambaut, A., Shapiro, B., and Pybus, O. G. (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22(5):1185–1192.

Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, (with discussion) 73(2):123–214.

Knorr-Held, L. and Rue, H. (2002). On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 29(4):597–614.

Minin, V. N., Bloomquist, E. W., and Suchard, M. A. (2008). Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution*, 25(7):1459–1471.

Pakman, A. and Paninski, L. (2014). Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *Journal of Computational and Graphical Statistics*, 23(2):518–542.

Shahbaba, B., Lan, S., Johnson, W. O., and Neal, R. (2013). Split Hamiltonian Monte Carlo. *Statistics and Computing*, pages 1–11.